Exploring Factors Affecting The Diagnosis of Alzheimer's Disease: A Statistical Machine Learning Approach

Aubree Krager, Lisette Villa, Nicholas Noel, Rasul Ibragimov, Ixhua Ramirez, Ruslan Ibragimov, Nkosi Sampson, Alex Jun, Lyla Traylor, Jared Brown, Andrew Young
Advisors: Dr. Sam Behseta, Dr. Archana McElligot, Dr. Mansour Abdoli

Abstract

In this study, we analyze whether demographic and health factors, such as race, ethnicity, sex, hypertension, diabetes, etc., affect the prevalence rates of Alzheimer's disease diagnoses using data provided by the National Alzheimer's Coordinating Center (NACC). Specifically, utilizing generalized linear models in conjunction with random forests, we find that some of these factors are significant. Subsequently, utilizing those factors identified in our models, in a neural network with one hidden layer allowed us to predict Alzheimer's Disease diagnosis with high accuracy. To facilitate this, we created a specific R package for NACC data sets and optimize current and future research goals.

Background Information and Goals

- In the United States, about 6.7 million adults age 65 and older live with Alzheimer's dementia
- Fifth leading cause of death for that age cohort [2]
- Characterized by Amyloid plaques, neurofibrillary tangles, and chronic inflammation
- Still, it is unclear if these are a cause or result of Alzheimer's disease [5].

Goals

In this study, we aim to address two primary goals. First, we aim to build an inferential model that allows us to gauge and assess risk factors associated with a higher prevalence of Alzheimer's disease diagnosis within the context of this data repository. Second, using those identified risk factors, we aim to create a reliable predictive model for Alzheimer's disease diagnosis.

Data Overview

The data set was fully de-identified and provided by NACC, a centralized data repository hub for 37 research and exploratory centers contributing data. Patients within the dataset may have more than one encounter with the care provider; this study primarily focuses on the most recent encounter chronologically from 2005-2022. Our study focused on participants' cognitive status and various health and demographic factors.

Data Structure and Strategies

Data Structure

- Uniform Data Set (UDS): A longitudinal data set with data concerning demographics, neurological examination results, diagnosis, and more.
- Biomarker Data Set (BDS): Data which contains summary measures for a subset of participants in the UDS.

Data Wrangling

- The data set was reformatted to convert missing and non-applicable values to NAs.
- We created an R package, **NACCdata**, that contributes to data cleaning and wrangling variables through multiple filters and tools.
- NACCdata streamlined exploratory data analysis and modeling, along with functions, this package is enough to make a poster in itself.

Literature Review

- A clinician diagnoses Alzheimer's by assessing a patient's cognitive symptoms; however, tools such as the Mini-Mental Status Exam have been shown to lack specificity and sensitivity among non-white individuals in addition to cultural differences and beliefs of the natural progression of age, spirituality, religion, and trust in healthcare providers. [1].
- Obesity is associated with a higher risk of Alzheimer's disease and damage to regions of the brain affected by it. However, a multitude of studies have found lower BMI is associated with a higher risk of progression to Alzheimer's disease [4].
- Studies have found smokers to have lower gray matter density and increased atrophy in the brain. Later in life, smokers present with similar neurocognitive issues to Alzheimer's and have been associated with an earlier onset of symptoms for those who do develop the disease. Decreased risk of Alzheimer's may be due to survival bias or exclusion from studies due to the development of other major diseases linked to smoking [3].

Methods

Modeling for building statistical inferences

A binomial logistic regression model was built with Alzheimer's Diagnosis as the response with the following two classes: normal cognition and Alzheimer's disease. The generalized model is written as follows for $i = 1, 2$ classes:

$$P(Y = i|X = x) = \frac{e^{xw_i+b_i}}{e^{xw_1+b_1+e^{xw_2+b_2}}}$$

- x - Observation
- b_i - Bias
- w_i - Weights

1. The logistic regression model used predictors obtained via preliminary data exploration and literature review which were then further subset through LASSO regression to maximize its predictive capacity.
2. A Random Forests™ model was also built using the same predictors and response found from the logistic regression. This model also included concentration of T-tau and $A\beta_{1-42}$ in the brain.
3. A Neural network with one hidden layer created a reliable predictive engine for Alzheimer's disease diagnosis with more than 83% accuracy

Multinomial Logistic Regression & Random Forests

Predictor
Depression
Alcohol Abuse
B12 Deficiency
History of Stroke
Hypertension
Hypercholesterolemia
Packs Smoked Per Day
Body Mass Index
Education (Years)

Table 1. Some significant variables. All variables had significant p-values.

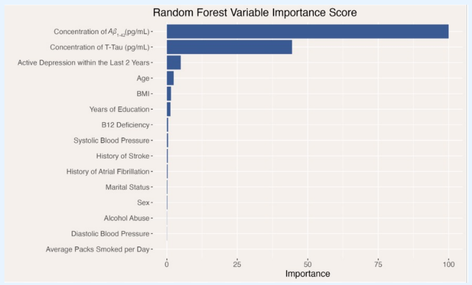

Figure 1. A variable importance plot from the Random Forests model.

Results

Comparative Analysis of Fitted Models

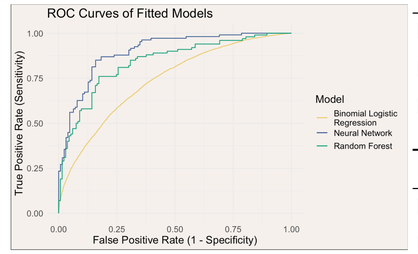

Figure 2. The receiver operating characteristic (ROC) curves of the following three fitted models.

Model Type	AUC
Random Forest	.85
Neural Network	.90
Logistic Regression	.73

Model Type	Accuracy
Random Forest	83%
Neural Network	83%
Logistic Regression	67 %

Table 2. A comparison of models by area under the ROC curve (AUC) and accuracy.

Above are the results from a binomial logistic regression built solely on the UDS data, along with a single hidden layer neural network and a random forest model, both built using variables from both the UDS and BDS data sets. For each, Alzheimer's disease diagnosis was the positive class, and normal cognition was the reference level. It is clear from table 2 above that the neural network and random forest methodology adhere to the highest accuracy rate in prediction. Note, that the Neural Network has a higher AUC, make it the better model. It is worth mentioning that this study was partially limited by a small number of counts in certain categories associated with a subset of variables, resulting in sparse data structures. To avoid that, we utilized filtering strategies to collapse categories with small counts.

Future Work

We would like to further investigate the relationship between demographic, environmental, and health factors and the progression and diagnosis of Alzheimer's disease, along with improving the accuracy of predicting diagnosis. We are currently implementing a linear mixed effects model using our longitudinal data and eventually build a Bayesian Hierarchical model.

Acknowledgments

This work was possible because of a generous donation from the Deland family, the LSAMP program, and the U-RISE Program. This work was also supported by various funding sources: U-RISE at CSUF NIH 5T34 GM149493-01, NICHD R03HD102448 and R01HD078547, and NSF (1826490).

References

[1] Selamawit Negash Ph.D. Roy Hamilton M.D. M.S. Alexander L. Chin, B.S. Diversity and disparity in dementia: The impact of ethnoracial differences in alzheimer's disease. *Alzheimer Disease Associated Disorders*, 25:187–195, 2011.

[2] Alzheimer's Association. 2023 alzheimer's disease facts and figures. *Alzheimer's dementia: the journal of the Alzheimer's Association*, 19,4:1598–1695, 2023.

[3] Timothy C. Durazzo, Niklas Mattsson, Michael W. Weiner, and Alzheimer's Disease Neuroimaging Initiative. Smoking and increased alzheimer's disease risk: A review of potential mechanisms. *Alzheimer's & Dementia*, 10(3S):S122–S145, 2014.

[4] Jena N Moody, Kate E Valerio, Alexander N Hasselbach, Sarah Prieto, Mark W Logue, and Hayes. Body Mass Index and Polygenic Risk for Alzheimer's Disease Predict Conversion to Alzheimer's Disease. *The Journals of Gerontology: Series A*, 76(8):1415–1422, 04 2021.

[5] National Institute on Aging [Internet]. What happens to the brain in alzheimer's disease?